Lecture 1 (4/3/17)

TIM 245 : Data Mining

Instructor : Tyler Munger

Agenda :

1) what is data mining

2) overview of the course

3) workload for the course

4) Brainstorming Exercise

① What is Data Mining

Data Mining: using data to Solve problems by answering questions

Two basic kinds of questions:

Descriptive Questions (unsupervised Learning)

Discovering useful (interesting) patterns in given input data X

- Cluster Analysis (groupings)
- Association Analysis (co-occurence)

Predictive Questions (Supervised Learning)

Build a model that can predict the output y given input x and historical training data (x, y)

- Prediction (numerical y)
- Classification (categorical y)

Example

Problem : Highway 17 is extremely dangerous, especially when the road is wet

Questions:

1) Which groups of drivers are most likely to have an accident?

We can answer this descriptive question using cluster analysis

Methods: k-means, hierarchical, DB-Scan

2) How many accidents are likely to occur on a particular day?

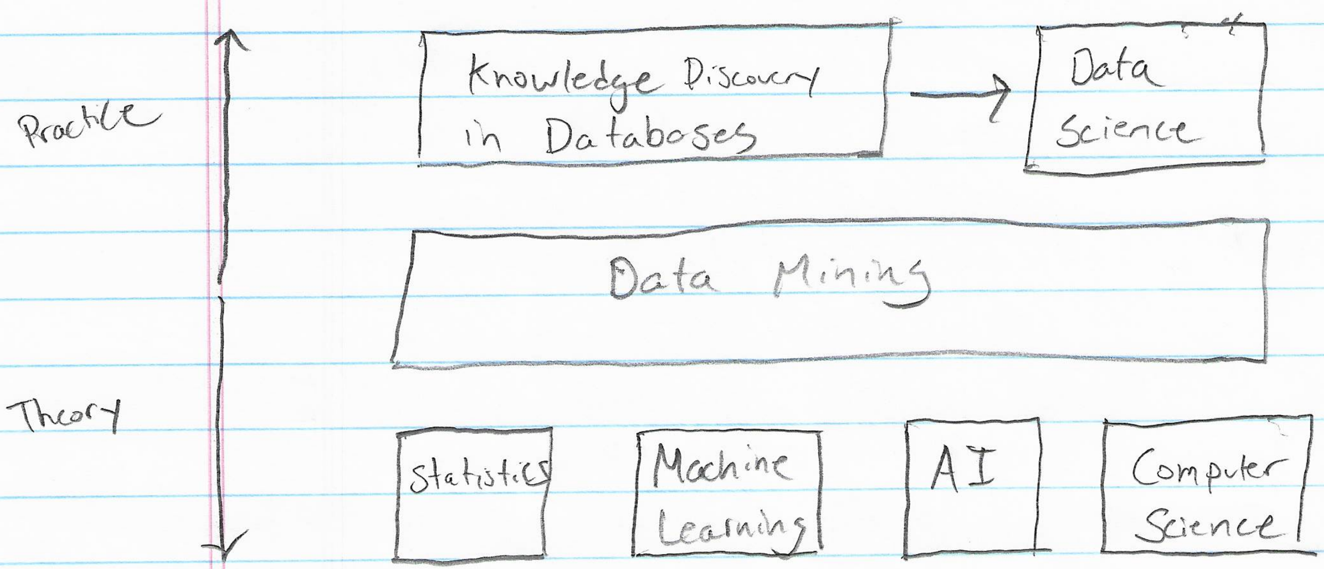
We can answer this predictive question using regression models.

Methods: Linear regression, Ridge/Lasso, Regression trees, Time series analysis

② Overview of the Course

The objective of the course is to provide the building blocks (methods) and a framework (or methodology) for apply these methods to real world problems.

This objective requires covering a mix of data mining theory and practice.



③ Workbad for the Course

Every Wednesday you will be submitting either a HW set or a project report

Homework

Wednesday → Receive HW Set → Monday → discuss HW in class → Wednesday → Turn-in Completed HW

← 1 week →

Project

Wednesday → Receive project Assignment → Wednesday → Submit project report → Monday → Project Review with instructor

(4) Brainstorming Exercise

Data mining is frequently a generative process that involves brainstorming new ideas for solving the problem under consideration

One method: Structured Brainstorming (osborn ~1940s)

Step 1: Generate ideas to solve a particular problem. (ideas are descriptive or predictive data mining questions)

Problem: Highway 17 is extremely dangerous especially when the road is wet

Work in groups to generate a mix of 20-30 ideas for improving safety on Highway 17

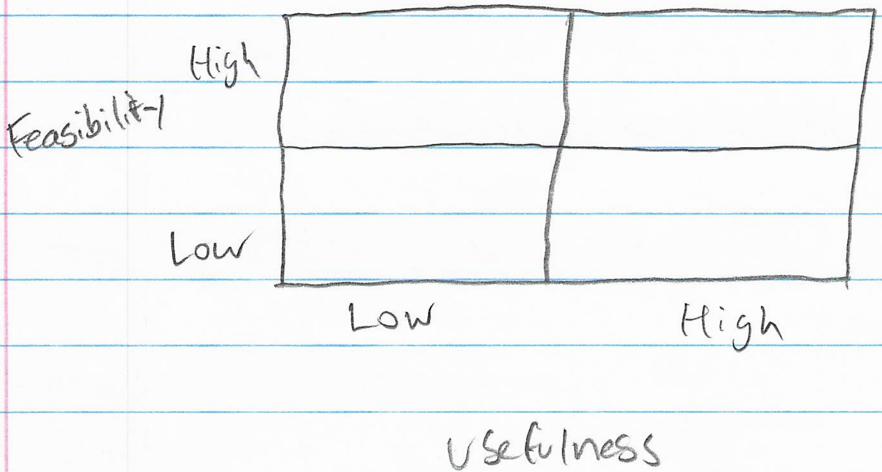
- Prediction
- Classification
- Clustering
- Association Analysis

Step 2: Structure the ideas into 3 groups

1) ideas that are immediately useful ("low hanging fruit")
e.g. predicting the number of accidents

2) Idea for further exploration
e.g. Clustering drivers

3) Idea that are radically new approaches
e.g. Classifying drivers in real-time


High
Feasibility
Low
Low
High
Usefulness